M 10016

THE OPEN UNIVERSITY

Mathematics Foundation Course Unit 16

Probability and Statistics I

The Open University

Mathematics Foundation Course Unit 16

PROBABILITY AND STATISTICS I

Prepared by the Mathematics Foundation Course Team

Correspondence Text 16

The Open University Press

Open University courses provide a method of study for independent learners through an integrated teaching system including textual material, radio and television programmes and short residential courses. This text is one of a series that make up the correspondence element of the Mathematics Foundation Course.

The Open University's courses represent a new system of university level education. Much of the teaching material is still in a developmental stage. Courses and course materials are, therefore, kept continually under revision. It is intended to issue regular up-dating notes as and when the need arises, and new editions will be brought out when necessary.

Further information on Open University courses may be obtained from The Admissions Office, The Open University, P.O. Box 48, Bletchley, Buckinghamshire.

The Open University Press
Walton Hall, Bletchley, Bucks

First Published 1971

Printed in Great Britain by
J W Arrowsmith Ltd, Bristol 3

SBN 335 01015 6

Contents

Note

In this unit we ask you to carry out a simple experiment. It will be necessary for you to have a pack of cards and a die — and a friend to help you with the experiment.

Objectives

This unit has two main objectives: to introduce you to methods of displaying and summarizing tabulated data and to introduce you to the concept of probability.

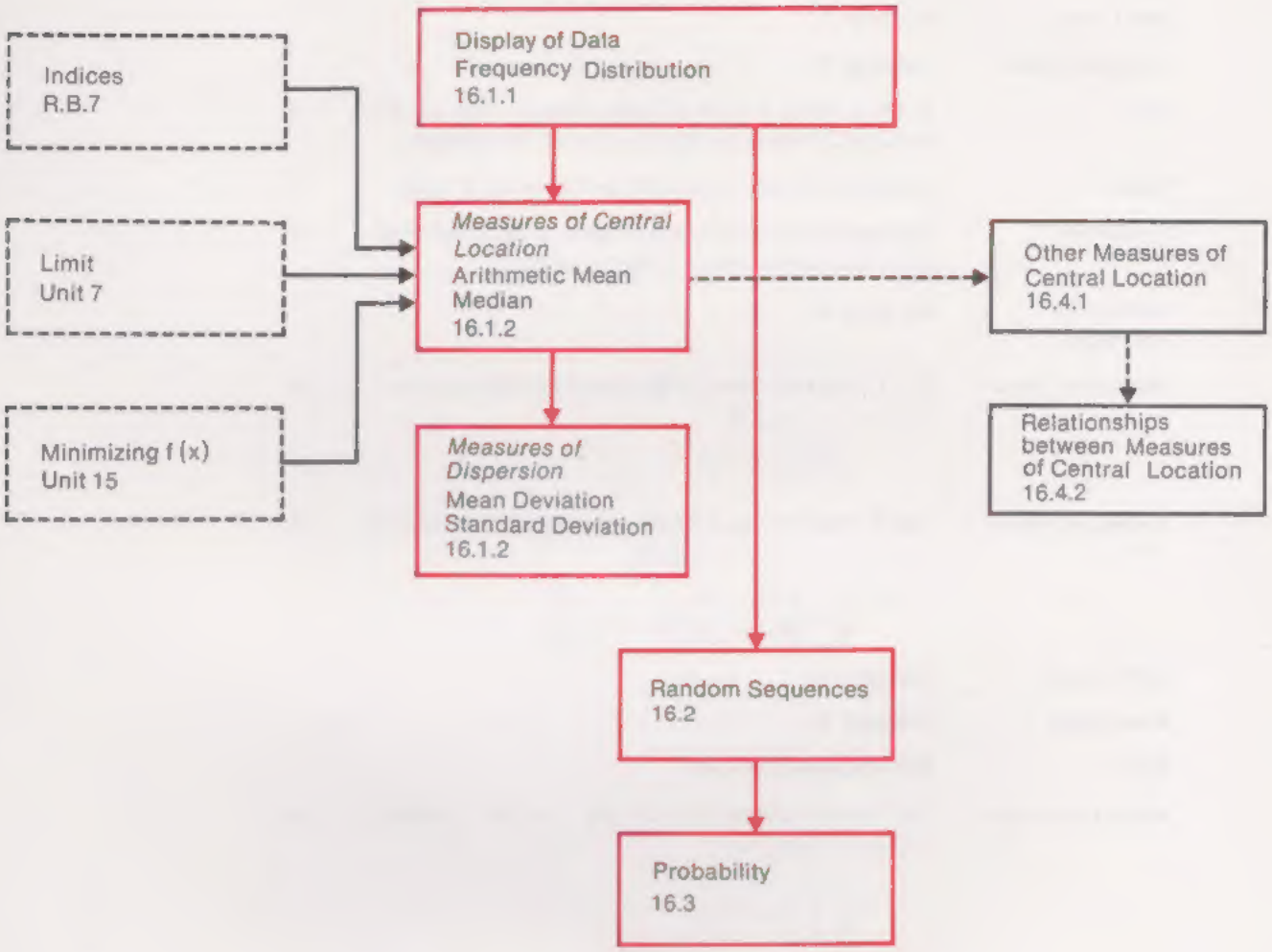
After working through this unit you should be able to:

(i) display a set of tabulated data in the most appropriate form;
(ii) explain the advantages and disadvantages of using bar charts, column charts, ideographs, pie charts, histograms and frequency polygons for the display of data;
(iii) point out how a given display of data may be liable to misinterpretation, and suggest ways of overcoming this;
(iv) explain in your own words the meanings of the terms:
mean,
median,
mean deviation,
standard deviation,
variance;
(v) calculate the mean, median, mean deviation and standard deviation of a given set of data, using "short cuts" where appropriate;
(vi) decide which statistics will be the most useful to summarize a particular set of data;
(vii) explain, as a result of personal experimentation, the meanings of the terms: random sequence, probability.

Note

Before working through this correspondence text, make sure you have read the general introduction to the mathematics course in the Study Guide, as this explains the philosophy underlying the whole course. You should also be familiar with the section which explains how a text is constructed and the meanings attached to the stars and other symbols in the margin, as this will help you to find your way through the text.

Structural Diagram



Glossary

Terms which are defined in this glossary are printed in CAPITALS.

		Page
ARITHMETIC MEAN	The ARITHMETIC MEAN (or just the MEAN) of the set of numbers $\{x_1, x_2, \ldots, x_n\}$ is $\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$.	17
BAR CHART	See page 2.	
COLUMN CHART	See page 8.	
DATA	A set of DATA is a set of items (which need not be numerical) which are used as a basis for inference.	1
EVENT	An EVENT is a set of possible outcomes of a TRIAL.	7
FREQUENCY	The FREQUENCY of an item of DATA is the number of times the item appears in the set of data.	2
FREQUENCY POLYGON	See page 16.	
GEOMETRIC MEAN	The GEOMETRIC MEAN of the set of positive numbers $\{x_1, x_2, \ldots, x_n\}$ is $\sqrt[n]{x_1 x_2 \ldots x_n}$.	35
HARMONIC MEAN	The HARMONIC MEAN of the set of positive numbers $\{x_1, x_2, \ldots, x_n\}$ is h where $\frac{1}{h} = \frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}\right)$.	35
HISTOGRAM	See page 10.	
IDEOGRAPH	See page 9.	
MEAN	See ARITHMETIC MEAN.	
MEAN DEVIATION	The MEAN DEVIATION of the set of numbers $\{x_1, x_2, \ldots, x_n\}$ is $\frac{1}{n}\sum_{i=1}^{n} \lvert x_i - m \rvert$ where m is the MEDIAN of $\{x_1, x_2, \ldots, x_n\}$. (Some authors take m to be the MEAN; see the discussion on page 24.)	24
MEASURE OF CENTRAL LOCATION	A MEASURE OF CENTRAL LOCATION is a measure indicating the "centre" of a set of numerical DATA; some examples are: the ARITHMETIC MEAN; the MEDIAN; the MODE.	17
MEASURE OF DISPERSION	A MEASURE OF DISPERSION is a measure indicating how numerical DATA are scattered about a "centre".	23
MEDIAN	The MEDIAN of a set of numbers with an odd number of elements is the number which occurs in the middle when they are arranged in order of magnitude. The MEDIAN of a set of numbers with an even number of elements is the ARITHMETIC MEAN of the two numbers which occur in the middle when they are arranged in order of magnitude. If $x_1, x_2, \ldots, x_{2n+1}$ are numbers arranged in order of magnitude, then the MEDIAN of $\{x_1, x_2, \ldots, x_{2n+1}\}$ is x_{n+1}, and the MEDIAN of $\{x_1, x_2, \ldots, x_{2n}\}$ is $\frac{1}{2}(x_n + x_{n+1})$.	21

		Page
MODE	The MODE of a set of numbers is the number which occurs most frequently in the set.	35
PIE CHART	See page 9.	
PROBABILITY	The PROBABILITY of an EVENT A is associated with the relative frequency of A in a sequence of TRIALS; it is not defined formally in this text.	34
RANDOM	A result which we have no means of predicting is said to be RANDOM.	30
RELATIVE FREQUENCY	The RELATIVE FREQUENCY of an EVENT E in a series of TRIALS is the ratio: $$\frac{\text{number of occurrences of } E}{\text{total number of performances of trial}}$$	31
RANGE	The RANGE of a set of numerical DATA is the difference between the largest number and the smallest number.	24
STANDARD DEVIATION	The STANDARD DEVIATION of a set of numbers is the positive square root of its VARIANCE.	26
TRIAL	A TRIAL is an experiment whose outcome need not be the same every time it is repeated.	34
VARIANCE	The VARIANCE of the set of numbers $\{x_1, x_2, \ldots, x_n\}$ is $$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$ where $\bar{x}$ is the ARITHMETIC MEAN of $\{x_1, x_2, \ldots, x_n\}$.	25
WEIGHTED MEAN	The WEIGHTED MEAN of the numbers $x_1, x_2, \ldots, x_n$, with weights $w_1, w_2, \ldots, w_n$ respectively, is $$\frac{w_1x_1 + w_2x_2 + \cdots + w_nx_n}{w_1 + w_2 + \cdots + w_n}.$$	20

Notation

Page

The symbols are presented in the order in which they appear in the text.

$\bar{x}$	The arithmetic mean of the set of numbers $\{x_1, x_2, \ldots, x_n\}$.	17
m_x	The median of the set of numbers $\{x_1, x_2, \ldots, x_n\}$.	21
s^2	The variance.	25
s	The standard deviation.	26
s_x	The standard deviation of the set of numbers $\{x_1, x_2, \ldots, x_n\}$.	26

Bibliography

M. Bruckheimer and A. Steward, *Index Numbers* (Chatto and Windus 1970).

This little book has a number of interesting pictorial representations taken from real life situations. Calculations of the mean, variance and standard deviation are clearly set out and illustrated by examples.

M. R. Spiegel, *Outline of Theory and Problems of Statistics* (Schaum 1961).

You will find this a useful book; it has many solved problems. The following chapters are relevant to this unit.

Chapters 1 and 2 deal with pictorial representations.

Chapter 3 discusses the arithmetic mean, median and the mode, also the relation between arithmetic and geometric means.

Chapter 4 discusses measures of dispersion, such as standard deviation and the variance.

Chapter 6 discusses elementary probability theory.

16.0 INTRODUCTION

16.0

Introduction

In the Mathematics Foundation Course there are three units on probability and statistics. Part of the general picture of mathematics and its applications at the present time must include something about these topics; indeed, a number of "O"- and "A"-level mathematics courses now include some probability and statistics. You may not, however, have met the subject before in any formal way. Partly because of this, and partly because it is the most logical procedure, we shall begin at the beginning

From the academic point of view, we shall not so much be giving a developed theory of probability as attempting to answer the question: "What is Probability?" At the end of this unit we introduce the concept of probability; in *Unit 18* we shall show how probability behaves, regarding it from a more theoretical stand-point. In *Unit 21* we shall attempt to answer the question: "What is Statistics?"

Because the chronological development of the subject has been so different from that of the rest of mathematics, it is worth having a brief look at the historical setting. Geometry was being developed and formalized by the Greeks 2000 years ago. Algebra, already in existence in Greek days, had by the sixteenth century advanced to the solution of cubic and quartic equations. Calculus was developed independently by Newton and Leibniz in about 1670, and mechanics together with gravitation theory also originated in the work of Newton. Complex numbers were recognized as early as 1550, and before the turn of the present century the theory of functions was well advanced. During all this time where were probability and statistics?

Briefly, the situation was this. In the time of Fermat and Pascal (seventeenth century), people began to apply mathematics to gambling, and started to lay down the first principles of probability theory. About the first person to devote himself to a systematic study of statistical questions was Karl Pearson. Not until the present century did people feel able to cope with an exact theory of small samples; the outstanding light at this stage was R. A. Fisher (an internationally famed statistician, who held the Chair of Genetics at Cambridge University). Yet despite even these advances, the first comprehensive paper on the theory behind significance testing (one of the most important ideas in statistics), by J. Neyman and E. S. Pearson, did not appear until 1936.

Since then progress has been tremendous. So much so, that it is now usual for experts to have their own specialities within the subject.

Karl Pearson 1857–1936

16.1 DATA

16.1

16.1.1 The Display of Data

16.1.1

Introduction

The word "statistics" has a number of meanings. More often than not it refers to the study of data; but a second meaning pertains to the data itself. Thus we say, "Have you seen the latest statistics?" or possibly, "What are her vital statistics?". Before we study data, we have to collect it; we then have to organize it and present it in a form displaying its distinctive features. We shall begin our study of statistics and probability by looking at some ways of presenting data.

Bar Charts

Example 1

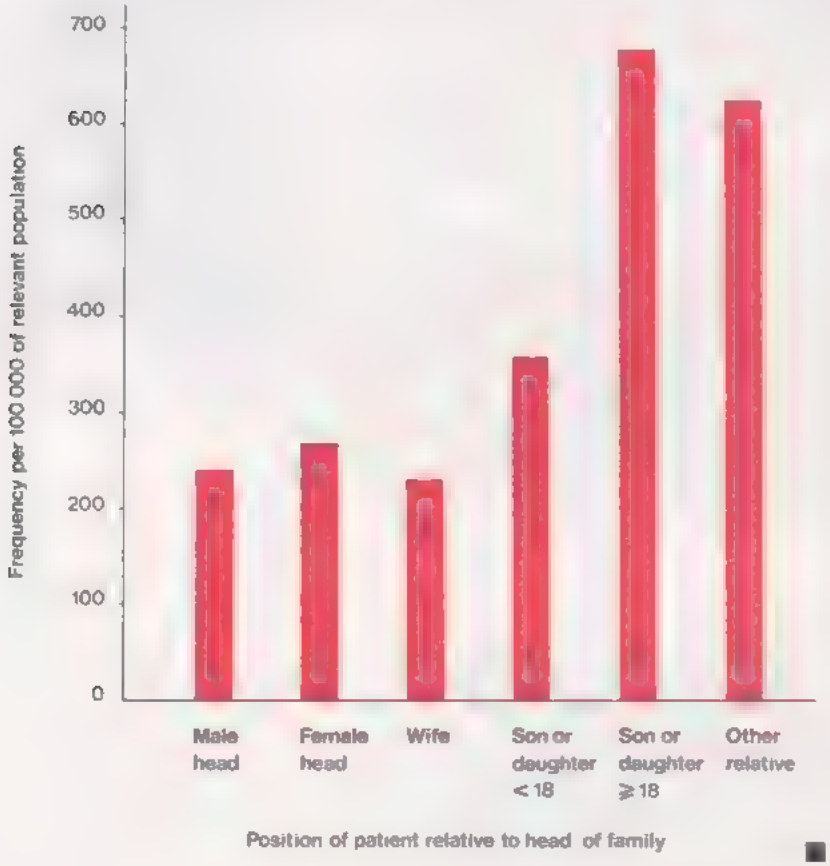
Example 1

First admissions to psychiatric wards in Maryland in 1960 have been analysed by the position of the patient in his or her family. The figures below are for patients belonging to households of 2 people only, and give the number of patients in each category per 100 000 of such households in Maryland. The number of patients in each category is called the frequency of that category.

Definition 1

Category of Patient	Frequency per 100 000
Male head of household	240
Female head of household	268
Wife of head	230
Son or daughter aged under 18	357
Son or daughter aged 18 or over	680
Other relative	624

The table shows that 0.24 % of 2-person families in Maryland had a male head who was a first admission in 1960, etc. The table itself is a way of displaying data, but it is only one of the ways of displaying this kind of information. We can also display the above information graphically or pictorially. Of course, a graphical or pictorial display cannot supply more information than the original table: in general there is a loss of information. Nevertheless, some graphs can give a visual emphasis of the significant points about the information. One type of graph used is known as the bar graph or the bar chart. The bar chart below represents the same information as that given in the above table but, as no doubt you will agree, it gives it in a more striking way.



■

It is important to label the two axes so that the reader can interpret easily the information being conveyed. Sometimes a condensed expression has to be used for the sake of brevity; if so, it should be clearly explained in the text in which the chart is embedded (viz. "relevant population"). The bars need not be actual rectangles as in the example: but they should obviously be conspicuous enough to catch the eye. The distance between the bars in this case has no significance, but for some data it may be naturally determined.

Exercise 1

Exercise 1
(3 minutes)

The figures below are for first admissions to psychiatric wards in Maryland in 1960 for patients belonging to households of 6 or more members.

Category of Patient	Frequency per 100 000
Male head of household	334
Female head of household	574
Wife of head	307
Son or daughter aged under 18	192
Son or daughter aged 18 or over	382
Other relative	249

Draw a bar chart for these data, and comment briefly on:

(i) distinctive features;
(ii) any changes from the "2-household" case. ■

Snags

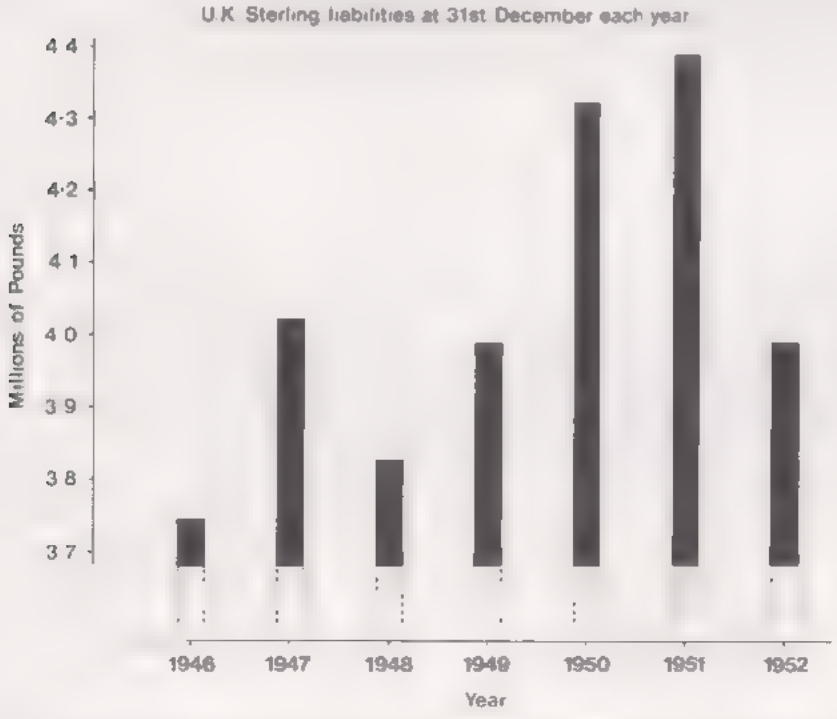
Discussion

Are there any pitfalls?

First, we consider a set of numbers which all lie close to some common large number; in this type of case authors often save space by not marking in the vertical axis in full

Example 2

Example 2

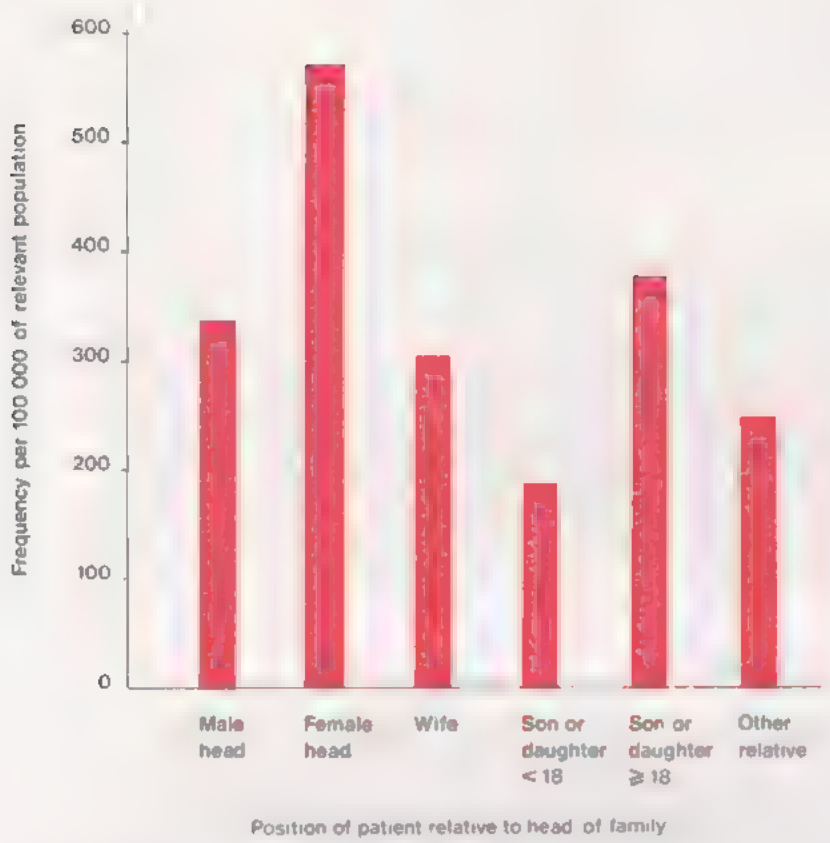


Note that broken lines are introduced to draw attention to the incomplete scale. ■

(continued on page 5)

Solution 1

Solution 1



(i) The obvious point is the predominance of female heads of the family. Sons and daughters under 18 are relatively small in number. Among the other categories there is not much variation.

(ii) The striking change is the increase for heads of the family, both male and female. Balancing this increase is the decrease for the dependants, whether children or other relatives. ■

There is nothing wrong with an incomplete scale so long as (i) the author is genuinely intending to save space rather than to deceive the eye, and (ii) the reader makes allowances. This second proviso is more easily stated than achieved; in some cases nothing short of redrawing the chart for oneself (if that is possible) will create a fair impression.

(continued from page 3)

Example 3

Example 3

An example of a chart which could convey misleading information is shown below. It is supposed to convey the same information as the chart in Example 2. But since the scale along the vertical axis and the broken lines have been deleted, no one could blame us for thinking, on the evidence of this chart, that (for example) the sterling liabilities in 1951 were about 8 times as great as in 1946. This conclusion is, of course, false. It is this sort of distortion in presentation of information against which we must guard, and we should always take care when interpreting such charts and graphs

■

Exercise 2

Exercise 2
(2 minutes)

If you were an unscrupulous managing director, would you use an incomplete scale to exhibit

(i) rising figures of sales?
(ii) falling sales?

■

Discussion

The spacings between the columns in the bar charts for sterling liabilities were (correctly) all the same, reflecting the fact that the events considered occurred at equal time intervals. The next example illustrates the false impression a bar chart can convey if the spacings between the bars do not properly represent the intervals involved.

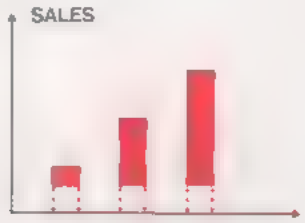
(continued on page 7)

Solution 2

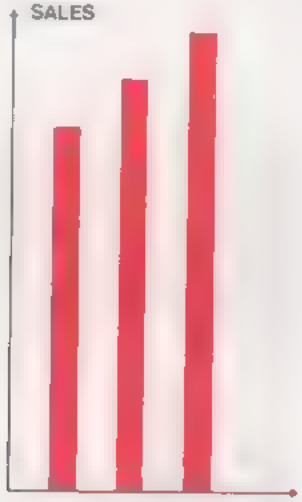
Solution 2

(i) Yes. As the changes are in the "right" direction, you would seek to give them as much emphasis as possible (unless perhaps you were dealing with tax authorities).

For example,



gives a much better impression than



(ii) No, for the complementary reason to the one given above. ■

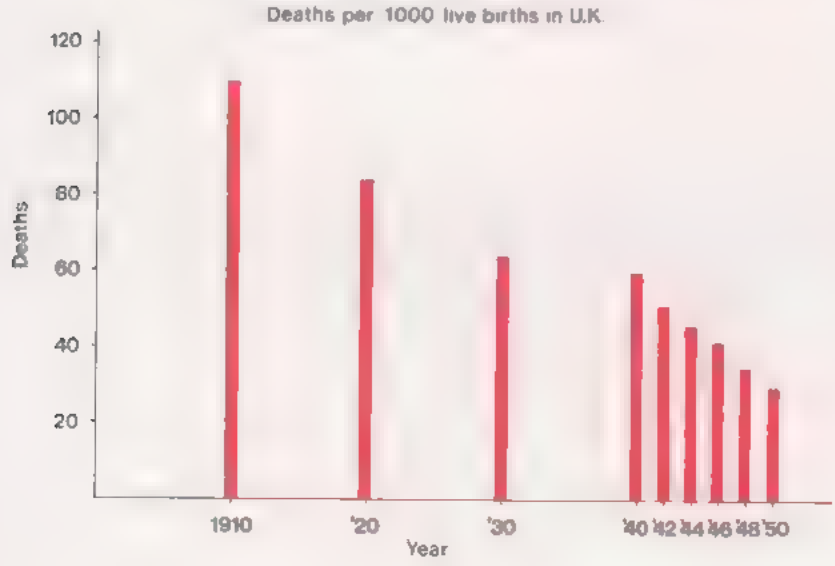
Example 4

Deaths of infants under 1 year of age per 1000 live births (U.K.)

1910	110
1920	82
1930	67
1940	60
1942	53
1944	48
1946	43
1948	36
1950	31

If we ignore the unequal spacing of the events when plotting the bar chart, we get the following diagram:

On the other hand, by spreading the years according to their chronological separations we get the chart below, which has quite a different visual effect.



■

Example 4

(continued from page 5)

Column Charts

If each plotted figure (or frequency) is the sum of a number of component figures, it may be necessary to show how the components are contributing towards the overall totals and variations. If we break up the bars into vertical columns (that is, rectangles), we can show different components by suitably shading or colouring different parts of the columns. We shall call such a figure a column chart. (No two authors, however, seem to agree on nomenclature.)

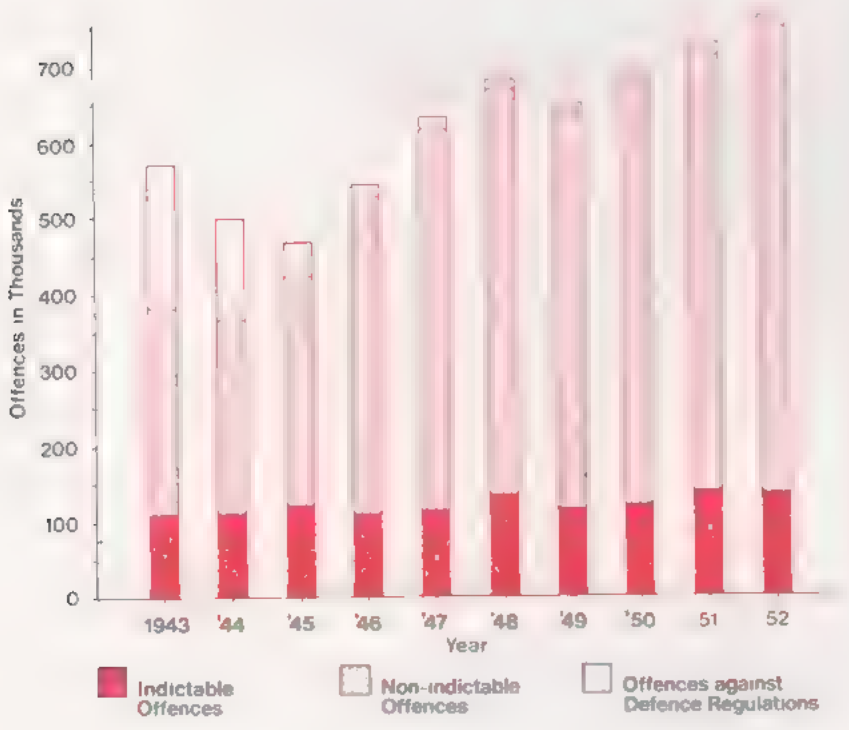
Example 5

Example 5

The chart below shows the number of persons found guilty of offences against the law in England and Wales, analysed by the nature of the offence. The figures in the table are given in thousands.

	Indictable Offences	Non-Indictable Offences	Offences Against Defence Regulations	All Offences
1943	105	278	187	570
1944	107	255	138	500
1945	116	298	54	458
1946	108	414	20	542
1947	116	498	19	633
1948	129	528	20	677
1949	114	523	13	650
1950	116	565	8	689
1951	133	584	6	723
1952	131	616	6	753

Plotted on a column chart, this gives:



■

Ideographs

Yet another way of presenting information is by using ideographs or pictograms. It consists of drawing pictures or caricatures of the objects involved.

Example 6 Example 6

The following table gives the annual output over 4 years of a factory manufacturing electronic valves.

Year	1947	1948	1949	1950
No. of valves	3000	4000	5500	7000

A corresponding ideograph is shown below:

■

We have discussed a number of different ways of displaying data: they all apparently use very little mathematics. They can, however, all be interpreted in terms of a basic mathematical concept — that of a *function*. For example, we were given a function which mapped each of certain years to the corresponding U.K. sterling liability. In each case, the domain of the function was either some subset of the set of real numbers (for example, year numbers) or a set of categories (for example, male heads of households); the codomain was a set of numbers: a set of frequencies. The choice we had in plotting the graphs of these functions was more a question of design than mathematics, although, as we saw, an improper choice of graph could lead to misrepresentation. Main Text

We shall now consider two other ways of displaying data, for which the data have to be processed.

Pie Charts

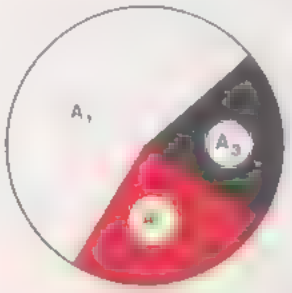
Pie chart presentations can be used when we consider a mapping whose domain contains a small number (generally less than eight) of elements. The frequency of each element (that is, its image under the mapping) is expressed as a percentage of the sum of the frequencies of all the elements in the domain, and this percentage is then represented by a sector of a circle. If one element of the domain occurs with a frequency of $Q\%$, then the angle of the corresponding sector is taken as $\frac{Q \times 360}{100}$ degrees, so that the arc length of the sector and the area of the sector both give visual representations of the frequency of that element.

Because we use a complete circle to represent the data, it is usual to use a pie chart only when the domain is "closed". For example, if in Example 6, the factory were still in business after 1950 (or before 1947), it would not be usual to use a pie chart. However, if it operated only for the four years for which the figures are given, then a pie chart could be used. The data of Example 1, on the other hand, is "closed", all categories of patients are considered, so a pie chart could be used.

Example 7

Example 7

A set of 100 coloured balls consists of a set, A_1, of 55 white balls, a set, A_2, of 25 red balls and a set, A_3, of 20 black balls. The corresponding pie chart is given below.



The area of each sector is proportional to the number of balls of the particular colour indicated

A_1 subtends an angle of 198° at the centre of the circle.
A_2 subtends an angle of 90° at the centre of the circle.
A_3 subtends an angle of 72° at the centre of the circle.

Histograms

In a sense a histogram may be considered as an extension of the pie chart, for it also uses areas to represent frequencies, but generally it necessitates choosing a suitable number of groups or categories into which the elements of the domain are partitioned. We represent each of these groups by a rectangle in such a way that the frequency which a rectangle represents is proportional to the area of the rectangle. In our first illustration we shall see that it is appropriate to choose all the rectangles to have the same width so that their areas are proportional to their heights; but in the second illustration we shall find that rectangles with different widths are appropriate.

Suppose we have the ages (at some reference date) of some 1 800 000 people. How do we represent the resulting age distribution? We *could* use the x-scale to represent the separate individuals and the y-scale to represent the ages. On this basis we would have 1 800 000 bars, but the bar chart so obtained would be confusing rather than informative.

One way out of the difficulty is to group the data. In the interest of clear presentation we simplify our description of the data and abandon the attempt to display *every* individual's exact age. Instead, we divide the range of possible ages into intervals: say under 5 years, between 5 and 10 years, and so on, and consider only the *group* to which each individual's age belongs. We shall need some convention for those people who fall exactly at the end of an interval, for example, someone who was 5 years old on the very day the ages were recorded. When that has been settled, we can use the bar chart *method* to display this simplified version of the data completely, by showing the number of individuals in each group.

Instead of keeping the bars separated as before, we broaden each bar into a rectangle, and we use their edges to show the width of the grouping intervals used, and so neighbouring bars touch. As long as all the grouping intervals have equal widths, the areas of the bars are proportional to their heights and so the resulting bar chart is also a histogram as defined above

Example 8

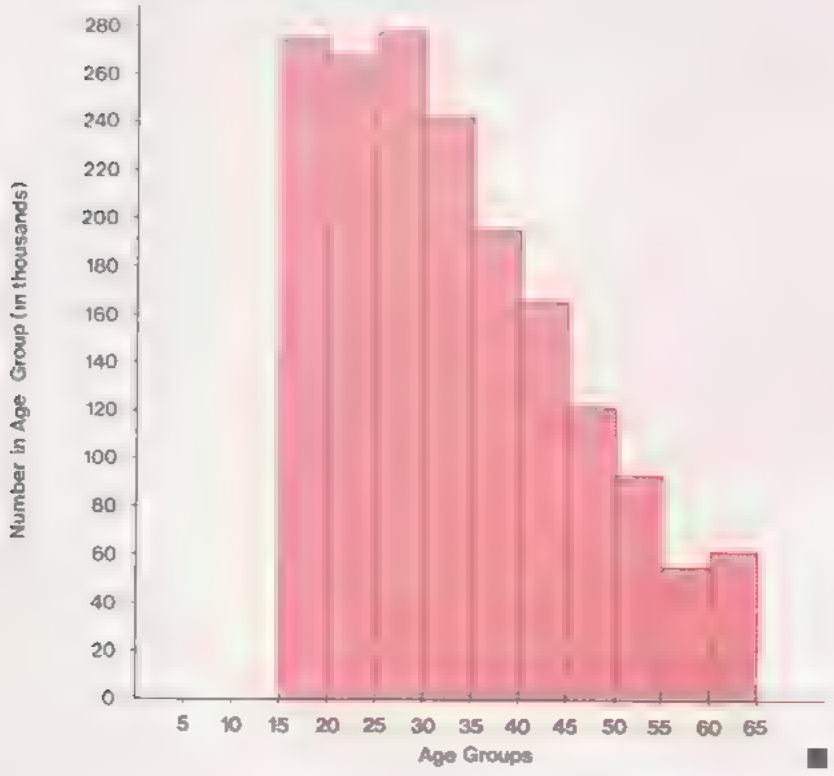
Example 8

A grouping such as that described above has been carried out for the male working population of Ghana in 1960 (ages 15 to 65), the figures being as follows:

Age	Number in Group to the nearest 1000 (in thousands)
15–20	276
20–25	268
25–30	279
30–35	243
35–40	198
40–45	166
45–50	123
50–55	97
55–60	59
60–65	63

In this case, anybody whose birthday fell on the day of recording, and whose age is at one of the end-points of the interval, has been put in the next group up. So, for example, anybody who was 25 on the date of recording would go into the 25–30 group. (Not that this is terribly important in this example, since the numbers have in any case been rounded to the nearest thousand.)

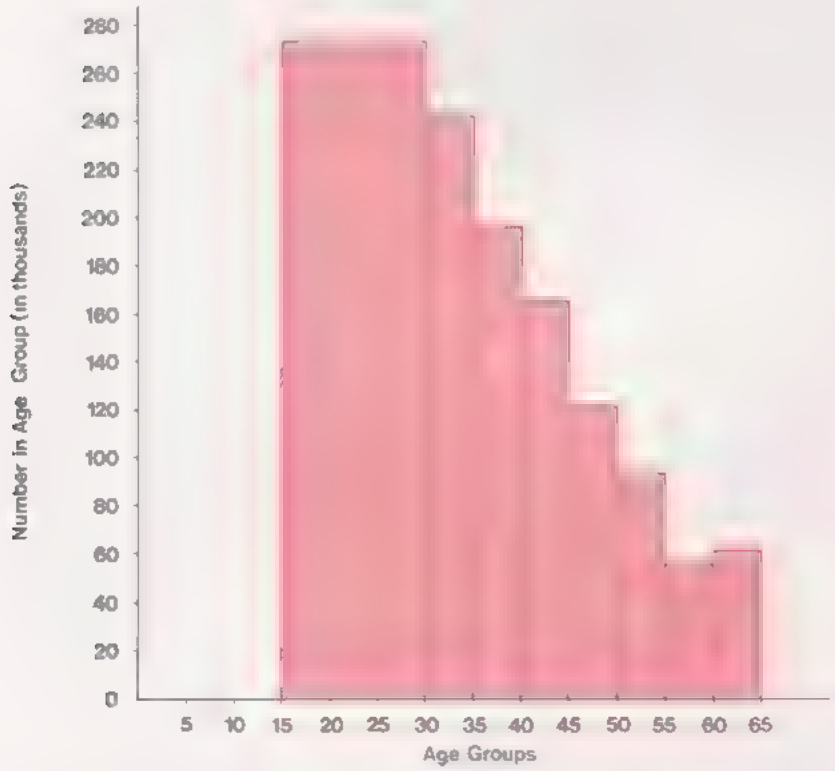
The corresponding histogram is shown below:



■

The essential difference between a histogram and a bar chart is that the former is used only when the data consists of numbers on a continuous scale (for example, the weights of ball bearings); then we have rectangles stretching from, say, a to b on the x-scale and representing the number of readings between a and b.

A representational difference between histograms and bar charts is that in a histogram it is the *area* of the rectangle which is proportional to the corresponding frequency, and in a bar chart the *height* of the bar is proportional to the frequency. We illustrate this point by looking again at Example 8. Suppose we combine the age groups 15–20, 20–25 and 25–30 into a single age group 15–30, and leave the others unchanged. The resulting frequency is 276 + 268 + 279 = 823, which would give a bar chart that went off the page if the scale of the above diagram were used, and if we used a rectangle of the same width as the others to represent this triple group. But this is not our reason for rejecting the bar chart. The point is that on compounding our three age groups, we get a large rectangle which has three times the base of one of the original rectangles. If we added the heights together we would get something like three times the height of one of the original rectangles as well, thus the area of the large rectangle would be roughly *nine* times that of one of the original rectangles. When comparing rectangles, as we do automatically here, it is the *area* (rather than height on its own) which conveys the impression of magnitude. For the area of the big rectangle to equal the sum of the areas of the three original rectangles (knowing that these smaller rectangles have equal bases) we must take a height which is not 823, but $\frac{823}{3}$, the average of the three separate heights. The resulting histogram is shown below.



It gives almost the same impression as before. The histogram is thus much less sensitive than the bar chart to the choice of grouping intervals, and can therefore be relied upon to give the intrinsic features of the data rather than those accidental features arising from the choice of intervals. In fact, in almost all cases of continuous-type data, the histogram gives a better presentation of data than a bar chart.

Sometimes the data tail off, and there may even be disconnected rectangles in a histogram. For example, if we take the age histogram of a small parish, there may be one person aged 102, and then no one else older than 93. In this case the rectangle corresponding to the person aged 102 will be isolated in the histogram. In five years' time he or she will probably have died, and the histogram on this occasion may consist entirely of connected rectangles, or the shape of its outline may have changed very little. As we are trying to exhibit *representative patterns* rather than chance effects, it would be more reasonable to group all data beyond a certain point into one band (wider than the others if necessary)

Exercise 3

Exercise 3
(3 minutes)

The following are hypothetical age figures for the retired inhabitants of a small English parish:

Age	Number
65–70	11
70–75	7
75–80	4
80–85	2
85–90	0
90–95	0
95–100	1

(i) Draw the histogram for the above data.

(ii) To avoid the separate rectangle on the right, draw the histogram again, taking the bands

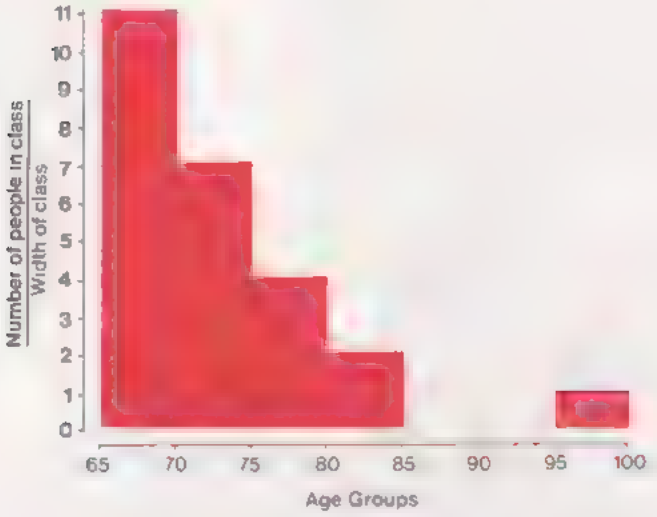
65–70
70–75
75–80
80–100

(iii) Suppose that, because of a slip or failing memory, the oldest inhabitant is not 96 as supposed, but only 94. It would now be possible to adopt a slightly different grouping to the one in (ii). What would an appropriate histogram look like in this case? ■

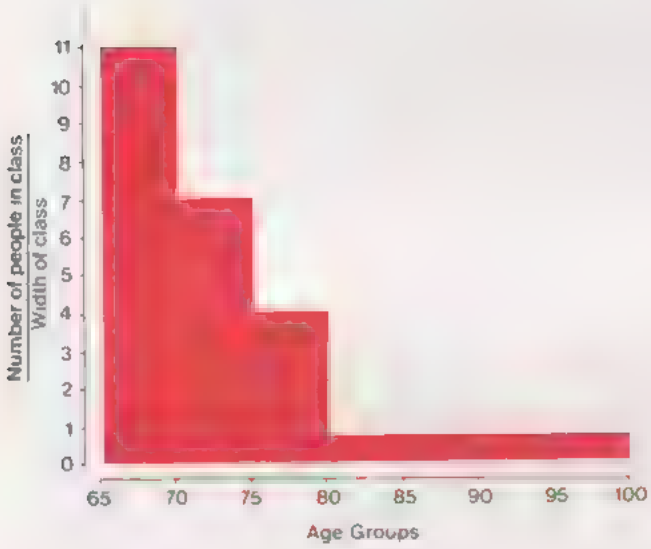
Solution 3

Solution 3

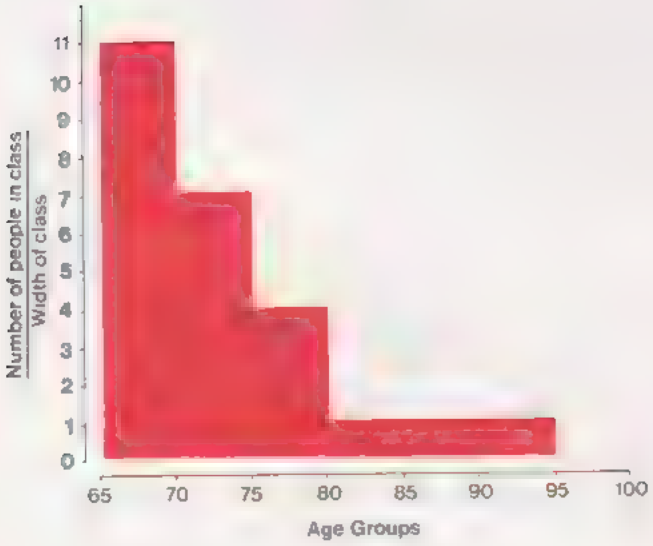
(i)



(ii)



(iii)

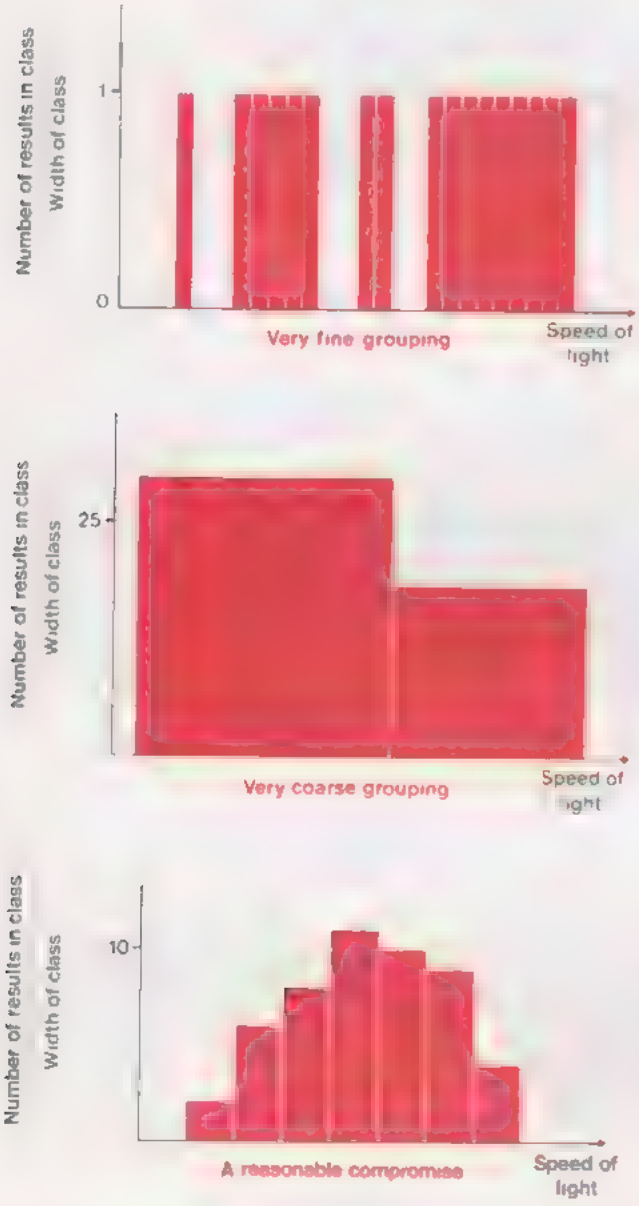


■

Discussion

We have seen that, before drawing a histogram, we have to decide on how to cope with isolated rectangles, if any. Either we can leave them separate, or we can combine them with others in wider rectangles. When the data are not already grouped for us, we also have to decide how to group them

Suppose a physicist carries out 50 times an experiment to measure the speed of light. Each time he gets a slightly different answer, and he wishes to draw a histogram of the results in order to examine the reliability of his experiment. If he took a very fine grouping, each interval would contain either 0 or 1 readings, and he would have 50 rectangles all of the same (non-zero) height. This would provide a pattern of density, but not of shape. At the other extreme, with a very wide grouping, he might have just one or two rectangles, which again would not give much idea of pattern. Between these two extremes of "too fine" and "too coarse" he hopes that there will be some reasonable compromise.



Sometimes we can decide how to draw an appropriate histogram just by looking at the data. Sometimes we can proceed only by trial and error; that is, draw one histogram, and if it looks inappropriate, scrap it and draw another

Frequency Polygons

One objection to the histogram is that the "shoulders" of the rectangles give the impression of discontinuities in the data. One way of removing these discontinuities in the pictorial representation is to mark in the centre points of the top sides of the various rectangles, and then to join these up by straight lines, thus forming part of a polygon. The figure thus obtained is called a frequency polygon; in this figure, the rectangles themselves will no longer appear

Exercise 4

Exercise 4
(2 minutes)

(i) Draw the frequency polygon corresponding to the data of Example 8.
(ii) Does the area under the polygon between 30 and 35 tell you anything about the frequency of ages between 30 and 35? ■

If the set of data is very extensive, and it is possible to take a very fine grouping (i e. very narrow band width), the frequency polygon may look more and more like an approximation to a curve. The associated histogram will then look very like the rectangles used in calculus leading to the definition of the area under a curve. The difference is that in calculus one proceeds from the curve to the rectangles, whereas here we are going the other way; we are inferring a curve from the rectangles.

16.1.2 Numerical Measures

16.1.2

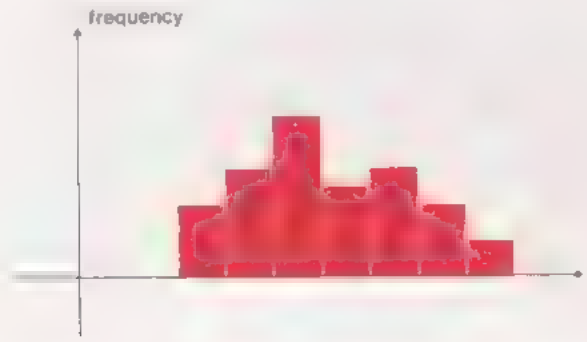
Introduction

Introduction

Given a set of statistical data, we have discussed different ways of displaying the relevant information. We sometimes wish to compare or contrast several sets of data. In order to do this, we could, for example, begin by constructing the corresponding histograms, but the problem remains: how do we compare or contrast the histograms in an objective way?

If possible, we would like to find a single number which would in some way represent *all* the data. Then, to compare, say, two sets of data, we would only need to compare these two representative numbers. In fact, however, even a crude representation of a data set necessitates *two* numbers

We can see this if, for example, we represent the data set by a histogram, say the one shown below.



One number is required to give an indication of the position of the "centre" of the histogram (that is, of the data set), and another number is required to give a measure of the "scatter" about the centre. There are various ways of defining such numbers. In this section we shall consider only a few of them.

We first consider some ways of defining a number to represent the "centre" of a given set of numbers. The new numbers so defined are called measures of central location.

Definition 1

The Arithmetic Mean

Main Text

Given the set* of n numbers, $\{x_1, x_2, \ldots, x_n\}$, we define the arithmetic mean of the set of numbers to be the number $\bar{x}$, where

Definition 2

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x}{n}$$

$\bar{x}$ is also known just as the mean of the set $\{x_1, x_2, \ldots, x_n\}$. (This measure is often called the average.)

Example 1

Example 1

The wages of 20 women working in a departmental store are given in the following table:

Content of Wage Packet (£)	Number of Packets
5	1
6	3
7	8
8	4
9	2
11	1
12	1

If we denote these women by the numbers $1, 2, \ldots, 20$, and the pay of the nth woman by x_n, the definition of the arithmetic mean gives the average wage of the 20 women as

$$\bar{x} = £\frac{x_1 + x_2 + \cdots + x_{20}}{20}$$

$$= £\frac{1 \times 5 + 3 \times 6 + 8 \times 7 + 4 \times 8 + 2 \times 9 + 1 \times 11 + 1 \times 12}{20}$$

$$= £7\,60$$

■

Exercise 1

Exercise 1
(2 minutes)

Decide whether each of the following statements is true or false.

(i) Given a set of numbers, one of them must be equal to their arithmetic mean. TRUE FALSE

(ii) The arithmetic mean of a set of numbers must lie between the greatest and least of the numbers. TRUE FALSE

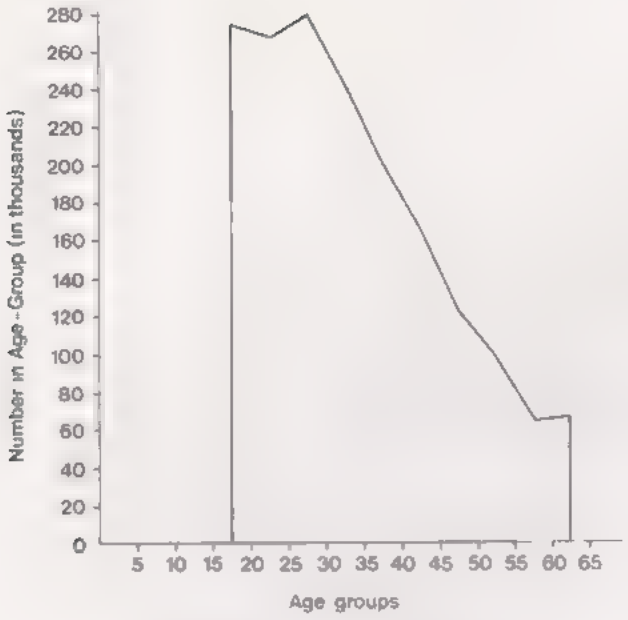
(iii) The arithmetic mean of a set of numbers must lie nearer to the centre of the range covered than to the extremities. TRUE FALSE

■

* In *Unit 1, Functions* we noted that, when listing the elements of a set, we do not record repetitions "Repetitions" in that context really means "unnecessary repetitions", that is, repetitions which do not carry additional information. In the statistical context, however, an identical measurement recorded twice *does* carry additional information, and therefore the x_i may not all be different

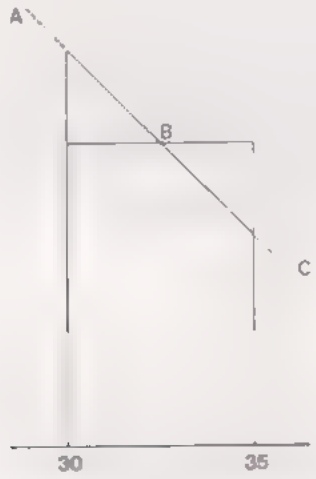
Solution 16.1.1.4 **Solution 16.1.1.4**

(i)



Frequency Polygon of Ages of Male Ghanaians (Working Population)

(ii) Yes. Let A, B, C be the vertices of the frequency polygon corresponding to the bands 25–30, 30–35, 35–40 respectively. A, B and C lie almost in a straight line; therefore the area under the polygon between 30 and 35 is almost the same as that of the rectangle whose base runs from 30 to 35 and whose upper edge passes through B.



This rectangle forms a component of the original histogram: therefore its area represents (on some scale) the frequency in the 30–35 age band. Therefore the given area under the frequency polygon has an area which approximately represents the same frequency. ■

Solution 1 **Solution 1**

(i) FALSE. The mean of the set of numbers $\{0, 2\}$ is 1, which is not a member of the set.

(ii) TRUE. Consider the set of numbers $\{x_1, \ldots, x_n\}$ where

$$x_1 \leqslant x_2 \leqslant \quad \leqslant x_n.$$

then

$$nx_1 \leqslant \sum_{i=1}^{n} x_i \leqslant nx_n,$$

hence

$$\frac{1}{n} \times nx_1 \leqslant \bar{x} \leqslant \frac{1}{n} \times nx_n$$

i.e.

$$x_1 \leqslant \bar{x} \leqslant x_n.$$

(iii) FALSE The mean of the set of numbers $\{10, 10, 10, 10, 1\}$ is $\frac{41}{5} = 8.2$, which is closer to 10 than to 5.5, the centre of the range. ■

Exercise 2

Exercise 2
(2 minutes)

Let $\{x_1, \ldots, x_n\}$, $\{y_1, \ldots, y_n\}$ be two sets of numbers with means denoted by $\bar{x}$ and $\bar{y}$ respectively.

(i) If $y_i = ax_i + b$ $(i = 1, \ldots, n)$ where a and b are real numbers, what is the expression for $\bar{y}$ in terms of $\bar{x}$?

(ii) If none of the x_i is zero and $y_i = \dfrac{1}{x_i}$ $(i = 1, \ldots, n)$, does it necessarily follow that $\bar{y} = \dfrac{1}{\bar{x}}$? ■

The result obtained in Exercise 2(i) is particularly useful, because it can often be of considerable help in numerical calculations. We illustrate this by taking three examples.

Example 2

Example 2

Suppose we require the mean of the set $\{101, 99, 104\}$. We can write

$$y_1 = 101 = \quad 1 + 100 = x_1 + 100, \quad \text{where } x_1 = 1;$$

$$y_2 = \quad 99 = -1 + 100 = x_2 + 100, \quad \text{where } x_2 = -1;$$

$$y_3 = 104 = \quad 4 + 100 = x_3 + 100, \quad \text{where } x_3 = 4.$$

Using the equation $\bar{y} = a\bar{x} + b$, and putting $a = 1$ and $b = 100$, we obtain

$$\bar{y} = 1 \cdot \bar{x} + 100$$

$$= 100 + \frac{1 - 1 + 4}{3}$$

$$= 101\tfrac{1}{3}.$$ ■

Example 3

Example 3

Suppose we require the mean of the set $\{175, 75, 50\}$.

We write

$$y_1 = 175 = 25 \times 7 = 25x_1, \quad \text{where } x_1 = 7;$$

$$y_2 = \quad 75 = 25 \times 3 = 25x_2, \quad \text{where } x_2 = 3;$$

$$y_3 = \quad 50 = 25 \times 2 = 25x_3, \quad \text{where } x_3 = 2.$$

Using the equation $\bar{y} = a\bar{x} + b$, and putting $a = 25$ and $b = 0$, we obtain $\bar{y} = 25\bar{x}$: here $\bar{x} = 4$ and so $\bar{y} = 25 \times 4 = 100$. ■

(continued on page 20)

Solution 2

Solution 2

(i) $\bar{y} = \frac{1}{n}(y_1 + y_2 + \cdots + y_n)$

$$\frac{1}{n}((ax_1 + b) + (ax_2 + b) + \cdots + (ax_n + b))$$

$$= \frac{a}{n}(x_1 + x_2 + \cdots + x_n) + \frac{nb}{n}$$

so $\bar{y} = a\bar{x} + b$.

(ii) No. For example, if $x_1 = 2$, $x_2 = 4$, then $\bar{x} = 3$, but $y_1 = \frac{1}{2}$, $y_2 = \frac{1}{4}$ and $\bar{y} = \frac{3}{8} \neq \frac{1}{3}$. ■

(continued from page 19)

Example 4

Example 4

We can calculate the mean of the set $\{5050, 5075, 5175\}$ by writing

$$y_1 = 5050 = 25x_1 + 5000, \quad \text{where } x_1 = 2;$$

$$y_2 = 5075 = 25x_2 + 5000, \quad \text{where } x_2 = 3;$$

$$y_3 = 5175 = 25x_3 + 5000, \quad \text{where } x_3 = 7.$$

Using the equation $\bar{y} = a\bar{x} + b$, and putting $a = 25$ and $b = 5000$, we obtain $\bar{y} = 25\bar{x} + 5000$

That is, we find $\bar{y}$ by evaluating the *simpler* mean $\bar{x}$: here $\bar{x} = 4$ and so $\bar{y} = 5100$. ■

Discussion

One is often confronted with the problem of combining two sets of data of a very similar nature. For example, we might know that 30 candidates took an examination consisting of two papers, and we may know the mean mark on Paper I and the mean mark on Paper II. Because of the way that the mean has been defined, we know immediately that the mean *total* mark is obtained just by adding these two means

In general terms, if the set $\{x_1, x_2, \ldots, x_n\}$ has mean $\bar{x}$ and the set $\{y_1, y_2, \ldots, y_n\}$ has mean $\bar{y}$, then the mean of the set

$$\{x_1 + x_1, x_2 + y_2, \ldots, x_n + y_n\}$$

is $\bar{x} + \bar{y}$.

That is,

the mean of the "sum" = the sum of the means.

In formal terms, let S be the set of all sets consisting of n numbers, and define the "sum" of two sets in S as above. Then the mapping

$$\{x_1, \ldots, x_n\} \longmapsto \bar{x}$$

is a *morphism* of $(S, \text{"sum"})$ to $(R, +)$. As different sets can have the same mean, this mapping is a homorphism.

Exercise 3

Exercise 3
(2 minutes)

If a set of m numbers has mean $\bar{x}$ and another set of n numbers has mean $\bar{y}$, calculate the mean of the combined set of $m + n$ numbers. ■

Main Text

The number $\frac{mx + ny}{m + n}$, where $x, y, m, n \in R$, is called the *weighted mean* of the numbers x and y with *weights* m and n respectively. In the general case, we define the weighted mean of the numbers $x_1, x_2, \ldots, x_n$, with

Definition 3

weights $w_1, w_2, \ldots, w_n$ respectively, to be the number

$$w_1x_1 + w_2x_2 + \ldots$$
$$w_1 + w_2 + \ldots + w_n$$

The weights, $w_1, w_2, \ldots, w_n$, are numbers which are not necessarily frequencies. (We have already found a weighted mean in Example 1; in that case, the numbers $x_1, x_2, \ldots, x_7$ were the contents of the wage packets, and the weights $w_1, w_2, \ldots, w_7$ were the corresponding numbers of packets.)

Median

Another measure of central location is the *median*, which is explained in the following illustration.

Suppose that on the board of a private company there is

one member earning £2000 per annum;
one member earning £2500 per annum;
one member earning £3000 per annum;
one member earning £3500 per annum;

and the chairman

earning £9000 per annum.

The mean salary is £4000 per annum. "Any company that can offer an average salary of £4000 to its board members is doing creditably." It *looks* as though it would be well worth trying to get onto the board.

The hard fact remains that, the chairman apart, no member achieves the mean salary of £4000, let alone exceeds it. Even including the chairman, the "middle member" gets only £3000, and any intending applicant is bound to pay more attention to the figure of £3000 than to the mean of £4000. We see that the consideration of *middle values* is useful in this sort of situation.

Given an *odd* number of numbers arranged in an increasing (or decreasing) order, there will be a middle number; this middle number is called the median of the set of numbers. For example, arranging the elements of $\{7, 6, 4, 6, 10\}$ in increasing order, we get 4, 6, 6, 7, 10, and we see that the median is 6. We define the median of a set containing an *even* number of numbers to be the arithmetic mean of the two middle values. For example, the median of the set $\{2, 3, 5, 6, 8, 9\}$ is $\frac{5 + 6}{2} = 5.5$.

Definition 4

Exercise 4

Exercise 4
2 minutes)

(i) Find the median of the set of numbers,

$$\{8.4, 4.7, 3.9, 9.3, 2.6, 4.1, 5.1, 4.7\}.$$

(ii) There is always some number in a set equal to the median of the set.

TRUE FALSE?

(iii) If $\{x_1, x_2, \ldots, x_n\}$ has median m_x, what is the median, m_y, of $\{y_1, y_2, \ldots, y_n\}$, where a and b are real numbers and $y_i = ax_i + b$, $(i = 1, 2, \ldots, n)$?

(iv) If $x_i > 0$, and $y_i = \frac{1}{x_i}$ $(i = 1, 2, \ldots, n)$, then

$$m \quad \frac{1}{m_x}$$

TRUE FALSE?

■

Solution 3 **Solution 3**

If $\{x_1, x_2, \ldots, x_m\}$ has mean $\bar{x}$ and $\{y_1, y_2, \ldots, y_n\}$ has mean $\bar{y}$, then $\{x_1, \ldots, x_m, y_1, \ldots, y_n\}$ has mean

$$\frac{(x_1 + \cdots + x_m) + (y_1 + \cdots + y_n)}{m + n} = \frac{m\bar{x} + n\bar{y}}{m + n}.$$ ■

Solution 4 **Solution 4**

(i) The two middle numbers are 4.7 and 4.7. Therefore the median is 4.7.

(ii) This is true for sets containing an odd number of elements, but otherwise it is, in general, false.

(iii) Suppose $x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n$. Using the rules for inequalities (see *Unit 6, Inequalities*), we have:

if

$$a \geqslant 0,$$

then

$$ax_1 \leqslant ax_2 \leqslant \cdots \leqslant ax_n,$$

and

$$ax_1 + b \leqslant ax_2 + b \leqslant \cdots \leqslant ax_n + b;$$

if

$$a < 0,$$

then

$$ax_1 \geqslant ax_2 \geqslant \cdots \geqslant ax_n,$$

and

$$ax_1 + b \geqslant ax_2 + b \geqslant \cdots \geqslant ax_n + b.$$

So either

$$y_1 \leqslant y_2 \leqslant \cdots \leqslant y_n$$

or

$$y_1 \geqslant y_2 \geqslant \cdots \geqslant y_n.$$

If n is odd, let

$$n = 2k + 1 \qquad (k \in Z^+),$$

then

$$m_x = x_{k+1}, \qquad m_y = y_{k+1} = am_x + b;$$

if n is even, let

$$n = 2k \qquad (k \in Z^+),$$

then

$$m_x = \frac{x_k + x_{k+1}}{2}, \qquad m_y = \frac{y_k + y_{k+1}}{2}$$

$$= \frac{ax_k + b + ax_{k+1} + b}{2}$$

$$= am_x + b.$$

Hence

$$m_y = am_x + b.$$

(iv) The statement is true if n is odd. Suppose n is odd and

$$x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n.$$

Since $x_i > 0$ $(i = 1, \ldots, n)$, it follows that

$$\frac{1}{x_1} \geqslant \frac{1}{x_2} \geqslant \cdots \geqslant \frac{1}{x_n}$$

(see *Unit 6, Inequalities*); that is, $y_1 \geqslant y_2 \geqslant \cdots \geqslant y_n$. So the middle y, m_y, corresponds to the middle x, m_x, and therefore $m_y = \dfrac{1}{m_x}$.

The statement is not generally true if n is even. For example, the median of $\{\frac{1}{2}, 2\}$ is $\frac{5}{4}$, and the median of $\{2, \frac{1}{2}\}$ is $\frac{5}{4}$, not $\frac{4}{5}$. ■

Discussion

There are other measures of central location, but they are less frequently used than the mean and the median. We discuss them (for those who may be interested) in the supplementary material at the end of this text.

The most widely used measure of central location is the arithmetic mean: it is simple to calculate and easily manipulated algebraically. There are also good technical (statistical) reasons why the arithmetic mean is useful; we shall discuss these in a later course.

The median is also useful, it is certainly simple to calculate and has similar algebraic properties to the mean (including the morphism property). We have seen an example in which the arithmetic mean was not a realistic indication of the "average" salary of a member on the board of a private company. The median gave a much better indication of the true state of affairs.

Measures of Dispersion

Main Text

In the introduction to this section we said that we would like to have *two* numbers to represent a set of data. We have called the number representing the centre *a measure of central location*, and we have given two examples, the *mean* and the *median*. We now have a look at numbers which measure "scatter". We call these measures of dispersion.

Definition 5

To discuss measures of dispersion of data about a central number, we begin by considering an example.

Example 5

Example 5

Hours of sunshine have been recorded in England and Wales, and the number of hours of sunshine in each month or year has been expressed as a percentage of some reference figure. The table below shows average percentages over February of each year, and also over each whole year.

Year	Annual Average (% of reference year)	February Average (% of reference February)
1943	107	129
1944	93	89
1945	96	99
1946	96	114
1947	99	47
1948	101	96
1949	117	159
1950	101	103
1951	100	96
1952	103	115

The means of the two columns of figures are fairly close; they are 101.3 for the annual figures and 104.7 for the February figures. Yet even without

plotting a bar chart, it is quite obvious that the two sets of figures are markedly different. In the first column the figures are relatively closely clustered, while in the second column they are spread out. This is clearly a distinguishing feature between the two cases, and so we would like to have a measure of the "scatter" in each case so that we can compare them. ■

Discussion

Measuring spread means measuring differences between numbers. Here we have only two basic options: we can measure either distances between the numbers themselves in some way, or else the separation of each number from some representative point. An example of the former measure is the range, which is defined to be the difference between the largest number and the smallest. However, the range is insensitive to what goes on between these two extremes, and so it is not a very good measure for the data as a whole. Another possible measure is the mean of the set of all numbers of the form $|x_i - x_j|$, where x_i and x_j both vary through all the recorded numbers, but $x_i \neq x_j$. However, it is not easy to interpret the numerical value of this measure in terms of the shape of the histogram. In any case, there are other and better measures, though what is meant here by "better" will have to be left to a later occasion (this is touched on in *Unit 21*).

Definition 6

A simpler procedure is to define a measure in terms of the separations of the numbers from a *fixed* number, a say, somewhere near the centre of the range:

$$|x_1 - a|, |x_2 - a|, \ldots, |x_n - a|.$$

For our measure of dispersion we could take the sum of these numbers or, more sensibly, their mean:

$$\frac{1}{n}\sum_{i=1}^{n} |x_i - a|.$$

If we take this expression, we shall obtain a quantity which varies with a. For a value of a which is very large and positive (or very large and negative) we shall obtain a very large measure. But this is ridiculous; we are trying to measure something about the data, not about a. Can we somehow eliminate a? Yes; we can do this if we choose a to depend in some way on the x's. We can take, for example, the value of a for which $\sum_{i=1}^{n} |x_i - a|$ is a minimum. The measure is then an intrinsic property of the x_i's, since it does not depend on any arbitrary number a.

Exercise 5

Exercise 5
(5 minutes)

(Don't spend too long on the first part read the solution if you have difficulty in getting started. Having seen the solution to the first part, you should be able to attempt the second part.)

(i) For the set of numbers $\{6, 7, 2, 2, 4, 10, 1\}$, what value of a minimizes $\sum_{i=1}^{7} |x_i - a|$, where x_i takes all the values in the set in turn?

HINT: Take $a < 1$, say -2; then consider what happens as a is increased

(ii) Generalize the above result for the set $\{x_1, x_2, \ldots, x_n\}$. ■

Discussion

If we want to minimize $\sum_{i=1}^{n} |x_i - a|$, then Exercise 5 suggests that we consider $\frac{1}{n}\sum_{i=1}^{n} |x_i - m_x|$, where a is chosen to be m_x, the median of the x's.

Definition 7

This expression is called the mean deviation from the median. Some people prefer to use the mean deviation from the mean, $\frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$, where a

is chosen to be $\bar{x}$. Although this does, of course, take into account the values of $x_1, x_2, \ldots, x_n$ in our choice of a, it does not satisfy our criterion of minimizing $\sum_{i=1}^{n} |x_i - a|$.

These measures are used occasionally, but they are rather awkward to manipulate because of the difficulty of taking the modulus of each term. Taking the modulus makes the deviation $|x_i - a|$ positive or zero for each i, and therefore the measure of dispersion is never negative. A more convenient way of obtaining such a measure is to square each of the terms in the sum, which gives $\frac{1}{n}\sum_{i=1}^{n} (x_i - a)^2$ as the measure of dispersion.

Once again we have a measure depending on a, and we can eliminate the effect of a by obtaining the minimum value. We can find this value of a in several ways; we shall describe two of them.

We wish to find the value of a which minimizes

$$F(a) = (x_1 - a)^2 + (x_2 - a)^2 + \cdots + (x_n - a)^2.$$

Using the methods of *Unit 15, Differentiation II*, we find that the stationary points of F occur when $F'(a) = 0$, that is, when

$$2(x_1 - a) + 2(x_2 - a) + \cdots + 2(x_n - a) = 0$$

or

$$(x_1 - a) + (x_2 - a) + \cdots + (x_n - a) = 0.$$

The required value of a is

$$\frac{x_1 + x_2 + \quad + x_n}{n} \quad x$$

The function F has an overall minimum at x

Alternatively, writing $F(a)$ in the form

$$F(a) = \sum_{\quad 1}^{n} (x_i - a)^2$$

$$= \sum_{\quad 1}^{n} x^2 \quad 2a \sum_{\quad}^{n} x \quad + na^2$$

we can collect all the a's in one term by adding $\frac{1}{n}\left(\sum_{i=1}^{n} x\right)^2$ to this expression

Of course, we must then subtract it again to balance things out. This gives

$$\sum_{\quad}^{n} x_i^2 + n\left(a \quad \frac{1}{n}\sum_{\quad}^{n} x\right)^2 \quad \frac{1}{n}\left(\sum_{\quad}^{n} x\right)^2$$

The only term containing a is the squared term in the middle. This will be a minimum when it is zero, that is, when

$$a \quad \frac{1}{n}\sum_{\quad}^{n} x \quad x$$

Thus, in this case, when we use the square, the "best" value to take for a is the mean, whereas in the other case, using the modulus, it was the median.

We therefore take as our measure of dispersion $\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$. We call this the variance

Definition 8

We acknowledge the presence of the square appearing in this measure by adopting the symbol s^2 for the variance. Thus

$$s^2 = \frac{1}{n}\sum_{\quad 1}^{n} (x_i - \bar{x})^2$$

(continued on page 26)

Solution 5

Solution 5

(i) Put the numbers on a line (we have put 2 both above and below the line to remind us that it occurs twice).

Begin with a less than all the given x's. Then

$$\sum_{i=1}^{7} |x_i - a| = \sum_{i=1}^{7} (x_i - a).$$

If a increases by $\frac{1}{2}$, and $a < 1$ still, then each of the seven terms decreases by $\frac{1}{2}$. Thus $\sum_{i=1}^{7} (x_i - a)$ decreases by $\frac{7}{2}$. This pattern continues as we increase a by $\frac{1}{2}$ until a reaches 1. At the next step, $|1 - a|$ increases by $\frac{1}{2}$, but the other six terms each decrease by $\frac{1}{2}$; on balance therefore there is a net decrease of $\frac{5}{2}$ in the sum. There will continue to be a net decrease with increasing a until there are as many x's to one side of a as to the other; that is, until a reaches the value 4.

(ii) Starting with a to the left of all the x's, $\sum_{i=1}^{n} |x_i - a|$ decreases until, if n is odd, a reaches the middle x, that is, the median. After that, $\sum_{1}^{n} |x_i - a|$ starts to increase. Thus the minimum occurs when a is the median. If n is even, $\sum^{n} |x_i - a|$ is constant and has its minimum value for all a between the middle two x's. Thus $\sum_{i=1}^{n} |x_i - a|$ still takes its minimum value when a is the median, but not only at this point. ■

(*continued from page 25*)

s, which is always taken to be positive, is called the standard deviation of the data. Thus the standard deviation is the square root of the variance.

Definition 9

Exercise 6

Exercise 6
(5 minutes)

If the set $\{x_1, x_2, \ldots, x_n\}$ has mean $\bar{x}$ and variance s_x^2, find the variance of $\{y_1, y_2, \ldots, y_n\}$, where:

(i) $y_i = ax_i$ $(i = 1, 2, \ldots, n)$ and a is a real number;
(ii) $y_i = x_i + b$ $(i = 1, 2, \ldots, n)$ and b is a real number;
(iii) $y_i = ax_i + b$ $(i = 1, 2, \ldots, n)$ and a and b are real numbers;
(iv) $y_i =$ $(i = 1, 2, \ldots, n)$. ■

When calculating a variance without a computer or a sophisticated calculating machine, it pays to use short cuts. The following identity gives a short cut which is often useful.

Discussion

Example 6

Example 6

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i^2 - 2\bar{x}x_i + \bar{x}^2)$$

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x} \sum_{i=1}^{n} x_i + n\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

■

This identity enables us to calculate the variance without having to find the actual deviations from the mean.

Exercise 7

Exercise 7
(3 minutes)

(i) For the sunshine data, given on page 23 (annual averages), would you square 107, 93, 96, etc to find the variance of the data?
(ii) What would you do first and why?
(iii) Find the variance for the data. ■

Summary

Summary

In conclusion, we note that if a is the median of the x's, it minimizes $\sum_{i=1}^{n} |(x_i - a)|$, and we may therefore think of this as being another property of the median considered as a central measure. Similarly, $a = \bar{x}$ minimizes $\sum_{i=1}^{n} (x_i - a)^2$, giving a property of the mean.

Solution 6

Solution 6

(i) If $\{x_1, x_2, \ldots, x_n\}$ has mean $\bar{x}$, then $\{y_1, y_2, \ldots, y_n\}$ has mean $a\bar{x}$. (See Exercise 4.) Therefore the variance of the y's is

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - a\bar{x})^2$$

$$\frac{1}{n}\sum_{i=1}^{n}(ax_i - a\bar{x})^2$$

$$= \frac{1}{n}a^2\sum_{i=1}^{n}(x - \bar{x})^2$$

$$= a^2 s_x^2.$$

This is the result we should have expected, because the y's are in effect the x's "spread out" by a factor a.

(ii) If $\{x_1, x_2, \ldots, x_n\}$ has mean $\bar{x}$, then $\{y_1, y_2, \ldots, y_n\}$ has mean $\bar{x} + b$. (See Exercise 4.) Therefore the variance of the y's is

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{x} - b)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i + b - \bar{x} - b)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= s_x^2$$

Thus, adding a constant to every term does not alter the variance; as we would expect, since this does not affect the spread.

(iii) We have shown, in (ii), that adding b to each term does not affect the variance; and in (i) we saw that multiplying each term by a has the effect of multiplying the variance by a^2. The net effect, then, is that the variance is $a^2 s_x^2$.

(iv) $$y_i = \frac{1}{s_x}x_i - \frac{\bar{x}}{s_x},$$

and so this is a special case of (iii) with $a = \frac{1}{s_x}$ and $b = -\frac{\bar{x}}{s_x}$. The variance is therefore

$$\left(\frac{1}{s_x}\right)^2 \times (\text{the variance of the } x\text{'s})$$

$$\left(\frac{1}{s_x}\right)^2 \times s_x^2$$

$$= 1.$$

■

Solution 7

Solution 7

(i) No.

(ii) Subtract 100 from each reading; this makes the numbers more manageable without affecting the variance (see Exercise 6(ii)).

(iii) Subtracting 100 from each value leaves:

x	x^2
7	49
−7	49
−4	16
−4	16
−1	1
1	1
17	289
1	1
0	0
3	9
+13	431

$$\sum_{i=1}^{10} x_i = 13$$

$$\left(\sum_{i=1}^{10} x_i\right)^2 = 169$$

$$\frac{1}{n}\left(\sum_{i=1}^{10} x_i\right)^2 = \frac{169}{10} = 16.9$$

Using the result of Example 6, we have:

$$\sum_{i=1}^{10} (x_i - \bar{x})^2 = \sum_{i=1}^{10} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{10} x_i\right)^2$$

$$= 431 - 16.9$$

$$= 414.1$$

$$\text{variance} = \frac{1}{10}\sum_{i=1}^{10} (x_i - \bar{x})^2$$

$$= 41.41$$

■

16.2 EXPERIMENT PRODUCING A RANDOM SEQUENCE

16.2

16.2.0 Introduction

16.2.0

Introduction

Many aircraft these days have automatic landing devices. Sometimes a fault develops and things go wrong; if so, the aircraft may get into difficulties. It may even crash. In order to keep our sense of perspective let us consider carefully what is really at stake.

If the automatic landing device goes wrong during the main flight, it does not matter, as the pilot should still be able to land the plane himself. If it goes wrong after touchdown, it clearly does not matter. In other words, the only time the plane is at risk is during a 30 second spell. No one is able to guarantee absolute safety in the air, any more than one can in trains or on the roads. The Air Registration Board (A.R.B.) is therefore prepared to tolerate a proportion of failures; but that proportion is as low as 1 in 10 000 000. To appreciate how low this is, suppose that landings had taken place every single hour, one per hour, day and night ever since the time of William the Conqueror, then, on average, we would be due for the first landing failure about now. In fact, while the failure rate allowed by the A.R.B. is 1 in 10 million, the design failure rate adopted by manufacturers in the U.K. is about 1 in 100 million.

The only *direct* way of seeing whether the risk tolerances are being met is to land 10^7 aircraft in succession (and even this would not be conclusive). Although we obviously cannot carry out such a mammoth experiment, we shall be very interested to know how the experimental results would behave, and to see what kind of predictions can be made about such a situation. After all, if there is no way of checking whether the A.R.B. condition is satisfied, it is a useless condition.

A result which we have no means of predicting is said to be random. The concept of randomness is fundamental to the theory of probability. As far as our aircraft is concerned, we shall record the digit 0 each time the aircraft lands safely and the digit 1 each time the aircraft has difficulties in landing due to failure of equipment. If the experiment is repeated many, many times we shall obtain a sequence of 0's, with possibly a few 1's interspersed. This is an example of a type of sequence which is so important in the theory of probability that we are going to suggest a number of simple ways in which such sequences can be obtained, and then consider some of their properties.

Definition 1

The best way to get some idea of this behaviour is to try out an experiment that shares with this hypothetical aircraft experiment the property that its results are apparently completely unpredictable. The conventional experiment of this type is to toss a penny repeatedly. If the result of the toss is "head", record a 1. If the result is "tail" record a 0.

16.2.1 An Experiment

16.2.1

Main Text

For a change from the conventional experiment, we suggest that you carry out a guessing game with a friend (or do your own guessing). This is the procedure:

Take a pack of playing cards, shuffle them, and then turn them face upwards one at a time. As each card is faced upwards, your friend, *who must not see the cards*, guesses the *suit*. If he (or she) guesses correctly, record a 1, otherwise a 0; record these numbers in the order in which they are obtained.

In this way you will build up a sequence of 52 0's and 1's. Now shuffle the pack and continue as before. Go on until you have 500 results (after

just under 10 deals of the pack). Record your results fairly neatly in a permanent form, because we shall refer back to them in later units of the Foundation Course.

There are now a number of investigations we would like you to carry out on your sequence.

Analysis of Results from the Experiment

Relative frequency of 1's

Divide your sequence up into blocks of 20 results: you will then have 25 blocks in all. Let

m_1 be the number of 1's in the first block,

m_2 be the number of 1's in the second block,

m_{25} be the number of 1's in the last block.

Now plot the points whose co-ordinates are

$$\left(20, \frac{m_1}{20}\right) = (20, \text{proportion of 1's in the first 20 results})$$

$$\left(40, \frac{m_1 + m_2}{40}\right) = (40, \text{proportion of 1's in the first 40 results}),$$

$$\left(60, \frac{m_1 + m_2 + m_3}{60}\right) = (60, \text{proportion of 1's in the first 60 results}),$$

and so on.

You will finish up with 25 points. These points are part of the graph of the function

$$n \longmapsto (\text{relative frequency of 1's in the first } n \text{ results}),$$

where by "relative frequency of 1's" we mean the proportion of times the result was a 1.

We expect the following statements about your points to be true (though remember that we have not seen them)

(i) As you progress along the sequence of points, they move up and down the page in an irregular manner
(ii) They never cease to be erratic, but
(iii) the further you move to the right, the smaller the irregular oscillations become
(iv) The points *appear* to settle down: that is, it *appears* that the sequence if continued indefinitely would have a limit.

The important issue is that the sequence does *appear* to have a limit. It is an extraordinary thing that these results, apparently completely irregular and unpredictable, in fact have this particular type of regularity which is so strong that we were able to predict the qualitative behaviour of your graph months before you drew it. In fact, it is possible to go even further than that and make some quantitative predictions, though with less confidence. If we had to predict your final point, we would say that it lay between the levels $y = 0.2$ and $y = 0.3$. If we were bolder, we would draw the boundaries of the interval in closer and say between 0.22 and 0.28. On the assumption that your results are not the rare exception (and there may be some very exceptional results among, say, 7,000 experiments), it is remarkable that we can make any kind of sweeping statement about a situation of this nature. One can only predict when there is a pattern or regularity. We have predicted. If we have predicted

successfully, that is empirical evidence that there is an underlying pattern and regularity of a sort, provided one looks in the right direction for it: and the right direction is towards the sequence of *relative frequencies*; that is, the sequence whose kth member is the number of 1's in the first k results, divided by k.

Skipping Terms

Go through the original sequence of individual 0's and 1's (we may refer to this as the 500-sequence) and form three sub-sequences as follows. Into sub-sequence A, put the elements whose positions are $1, 4, 7, 10, \ldots$, into B the elements whose positions are $2, 5, 8, 11, \ldots$ and into C the elements whose positions are $3, 6, 9, 12, \ldots$. Lastly, form a sub-sequence D as follows. Throw a die, and if (for example) a 4 comes up, put the 4th element of the sequence into D. Throw again, and if (for example) a 3 comes up, move along 3 places in the sequence and put this element (the 7th from the beginning) into D. Continue this process until you reach the end of the sequence. Now examine the final relative frequencies for A, B, C, D. Are they roughly equal to each other? If they all lie between 0.15 and 0.35, or at least between 0.1 and 0.4, they can be considered to be very nearly the same.

Runs

In coin tossing it is sometimes believed that a run of tails makes a head more likely next time. (This is sometimes called the "law of averages" but in fact it is a fallacy, not a law, as some gamblers have learnt to their cost.) To see whether there is any basis for this, you can turn once again to your empirical data. Go through the sequence again: and *whenever* you come across a run of three 0's in succession, write down the next number as an element of a *new* sequence. For example, if in the complete 500-sequence you have

$$1, 0, 0, 1, \underbrace{0, 0, 0}, 0, 1, 1, \underbrace{0, 0, 0}, 1, \underbrace{0, 0, 0}, \underbrace{0, 0}, 1, 0, 0, \ldots,$$

the digits you would mark down are those which appear in red. Notice that the same zero can appear in more than one run of three

Suppose you finish up with x digits in your new sequence. For this new sequence you can find the total relative frequency of 1's. Now divide up your 500-sequence into blocks of x; work out the relative frequency of 1's for each of these blocks separately (thereby getting a number of relative frequencies), and then decide whether the relative frequency for the new sequence is all that different from those for the blocks. If the occurrence of three successive 0's made a 1 more likely next time, the relative frequency in your new sequence should be noticeably greater than those in the blocks

Summary

Summary

The sequence of 0's and 1's you have produced, though at first sight entirely chaotic, has in fact a curious kind of regularity; its sequence of relative frequencies has the *appearance* of approaching a limit, and sub-sequences selected according to various rules also produce relative frequencies with about the same value. Such a sequence — completely unpredictable in the short run, but predictable in the long run — is called a random sequence. This concept of random sequence plays an important part in the concepts of probability and statistics.

Definition 2

We have taken a situation where chance and guess-work produce a randomness of behaviour which is unpredictable except as regards the value of the relative frequency of the occurrence of some situation. There are many other situations where we are faced with some inherently un-

predictable effect; for example, the decay of radioactive nuclei, the emission of X-rays, the incidence of disease, etc. There is a line of thought which suggests that these effects only appear to be unpredictable because we do not yet know enough about their mechanism. This may be true, but it is academic if, in fact, we do not know about their mechanism; in any case, even if we know the mechanism (as in the case of the spinning coin), this knowledge alone may not be sufficient for prediction. The only way in which we can analyse them is to treat them as random events, just like the card guessing results.

Exercise 1

Exercise 1
(2 minutes)

(i) Can you think of other situations where events behave as if by chance, even though a complete knowledge of the situation would allow you to work things out by inflexible cause and effect?
(ii) Can you think of any situation in which we rely on chance?
(iii) Can you think of any situation where we even strive to make things random on purpose? ■

Discussion

One of the drawbacks of randomness is that it is often a nuisance when we don't want it; but when we want it, it is difficult to achieve. In fact there are special tables of random numbers which are frequently used. You are probably aware that special machines are constructed to generate random numbers — ERNIE is an example: it chooses the numbers of winning premium bonds

Lastly, we mention some of those situations where randomness is the last thing we want; in fact, so little do we want it that we prefer to forget that it is there at all. Suppose a physicist is measuring the gravitational acceleration, g; as you may know, this can be done by timing the period of oscillation of a pendulum. If you carry out the experiment, you get an answer. If you do it again, you again get an answer. Are the two answers equal? Can they be expected to be equal?

Now admittedly a careless experimenter may well get a set of entirely different answers. So he must take greater care to get closer agreement. But will he ever be able to eliminate variability completely? Scarcely. The length of the suspension may depend on the temperature He controls the temperature; but can he keep it constant within an accuracy of 6 decimal places? Can he even measure it this accurately?

The period of oscillation depends *slightly* on the amplitude of swing; can this be controlled within an accuracy of 6 decimal places? The very swing itself may induce a minute "give" in the suspension. The timing mechanism must itself be accurate to a certain number of decimal places. It is difficult to know *exactly* when a full period has been completed. However we look at it, we must resign ourselves to a small element of uncontrollability; in other words, to randomness. There is no certainty for the manufacturer of aircraft components; there can be none for scientists measuring g. And compared with the finding of g, many "accurate" and "controlled" processes are crude indeed.

Does this mean there is nothing we can do about it? No! As soon as we admit that some degree of randomness is unavoidable, we can see a method of at least improving the situation. For, given a number of values all purporting to be values of g, we can select a suitable measure of central location. Intuitively, we feel that this will be a "better" estimate than a random choice of a single value from all those obtained.

Having explained, albeit in rather intuitive terms, what is meant by a random sequence, and having considered some of the properties of random sequences, we shall now attempt to explain what is meant by probability.

Solution 1

Solution 1

(i) The way in which sex and genetic characteristics are determined.
The ages at which people die (or are killed).
The moments in time when people initiate telephone calls.
The way gas molecules move around

(ii) We rely on the chance emission of neutrons to make the atom bomb work
One particular example of reliance on chance occurred at King's College, London, where parking space was limited. Dons were allowed to bring their cars in twice a week, and it was left to chance to spread the load evenly

(iii) We try to achieve randomness with ERNIE (the machine which selects premium bond winners), with the manufacture of dice, with the selection of personnel for sample survey questionnaires, etc. ■

16.3 PROBABILITY

16.3

Discussion

The card guessing experiment described in the previous section has (we hope) convinced you that

(i) the results of individual trials (suit guesses in our case) are unpredictable;
(ii) in an experiment consisting of a sequence of trials, and producing a sequence of 0's and 1's, the sequence is random;
(iii) nevertheless the sequence of relative frequencies of 1's *appears* to have a limit

(It must be clearly understood that this is an empirical result, and that "limit" is used in the everyday interpretation of the word rather than in the mathematical one — this is why *appears* is in italic.) The numerical value of this "limit" is called the *probability* of the particular result in question. For example, the card guessing experiment gives rise to a sequence of relative frequencies for correct guesses. This sequence *appears* to have a limit somewhere in the interval [0.2, 0.3]. We say that the probability of a correct guess appears to be in this interval (and we feel that we could locate it more accurately if we were to continue with a longer sequence of guesses).

This does not constitute a formal definition of probability, because it is too vague. It does not enable anyone to determine the exact value of the probability. Admittedly we can make the sequence of trials longer and longer, so that the relative frequency gets more and more stable; but this will never determine a limit, since we can only perform a *finite* number of trials. Indeed, it can only be an assumption on our part that the relative frequency would tend to a limit as the number of trials increased indefinitely. There is, however, a more fundamental objection.

In *Unit 7, Sequences and Limits 1*, a sequence $u_1, u_2, u_3, \ldots$ was said to tend to the limit p if for any small number $\varepsilon > 0$ all terms beyond the nth (for some n depending on ε) were *certain* to lie within the interval $[p - \varepsilon, p + \varepsilon]$. Moreover, this had to be true however small ε might be. For random sequences nothing is certain. However large n might be, it is always possible for the relative frequency after n trials to be outside the interval $[p - \varepsilon, p + \varepsilon]$: indeed, if it were impossible, the sequence in question would not be random.

While the relative frequency is not certain to lie within $[p - \varepsilon, p + \varepsilon]$ after n trials, it becomes more likely to do so as n increases. But we cannot rely on this to provide a definition of probability, because we have used

the phrase "more likely", and what does this mean other than "more probable"? We cannot say "if the probability of something increases with n, then we define probability to be ...": we are in a circular argument.

But while a formal definition has eluded us, we have acquired a good intuitive idea of relative frequency and of the way in which it becomes more stable as n increases. In an idealized sense, we can think of probability as being the value of the relative frequency in the limit. If in a practical finite case we have to take a plunge for some value or other, the odds are that we shall not be far wrong.

Postscript

"That some words have been explained and some hard ones left alone is more than likely, since, on such a subject, no standard exists either of information or of ignorance."

Augustine Birrell
Editor's note to
The Poetical Works
of Robert Browning

16.4 SUPPLEMENTARY MATERIAL

16.4

You may omit this material if you are short of time.

16.4.1 Other Measures of Central Location

16.4.1

Discussion

We have already mentioned the mean and median as measures of central location, because they are the two most important measures. In addition, however, there are the following measures:

(i) We can take the mid-point of the two extreme values.

(ii) We can take the most frequently occurring value or, if the data are represented by a histogram with equal group intervals, we can take the group which occurs the most frequently (or, more strictly, some suitable value within this group). This is called the mode. (Notice that there may be more than one mode.)

Definition 1

(iii) If the numbers $x_1, x_2, \ldots, x_n$ are all positive, we can take their geometric mean, which is

Definition 2

$$\sqrt[n]{x_1 x_2 \cdots x_n}$$

(iv) If the numbers $x_1, x_2, \ldots, x_n$ are all positive, we can also take the harmonic mean, h, which is defined by

Definition 3

$$\frac{1}{h} = \frac{1}{n}\left\{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}\right\}.$$

Let us consider each of these measures in turn.

The mid-point of the extremes has an obvious appeal; it comes bang in the middle of the range of the data, and so is a central point in the usual sense of the word. On the other hand, it does not indicate how the data lie between the two extremes. If it is important to know whether most numbers are at the lower end, for example, we obviously do not choose the mid-point of extremes as our measure of location.

The mode also has an obvious appeal, for it represents the "favourite" group. On the other hand, knowing the mode tells us nothing about groups away from the mode (except that each of them occurs relatively less frequently than the modal group).

The two measures above are in a way complementary. The former tells us something about the extreme values but nothing about the values in between; the latter tells us about the "favourite" group (usually somewhere in the middle — but not necessarily so) and very little about the groups at the edges.

The geometric mean at least depends on every single number, so it tells us something about the data as a whole; it is sometimes used by economists. However, it is greatly affected by the size of the smallest number if this is anywhere near zero, and must be used with care.

The harmonic mean of two numbers is sometimes used to interpolate certain tabulated functions, but it scarcely occurs otherwise.

Exercise 1

Exercise 1
(3 minutes)

Decide whether each of the following statements is true or false.

In each case assume that $x_i > 0$, $y_i > 0$ $(i = 1, 2, \ldots, n)$.

(i) If k is a real number and

$$y_i = x_i + k \qquad (i = 1, 2, \ldots, n),$$

then

(geometric mean of $y_1, y_2, \ldots, y_n$)
$= k +$ (geometric mean of $x_1, x_2, \ldots, x_n$). TRUE FALSE?

(ii) If a is a positive number and

$$y_i = ax_i \qquad (i = 1, 2, \ldots, n),$$

then

(geometric mean of $y_1, y_2, \ldots, y_n$)
$= a \times$ (geometric mean of $x_1, x_2, \ldots, x_n$). TRUE/FALSE?

(iii) If

$$y_i = \frac{1}{x_i} \qquad (i = 1, 2, \ldots, n),$$

then

(geometric mean of $y_1, y_2, \ldots, y_n$)
$= \dfrac{1}{\text{(geometric mean of } x_1, x_2, \ldots, x_n)}$. TRUE/FALSE?

■

16.4.2 Relationships Between the Measures of Central Location

Leaving the harmonic mean out of account, we have up to five different measures of central location. If we calculate them all for a given set of data, we shall, in general, get up to five different answers, so it would be sensible to begin by examining the relationship between them — if any.

First of all, the data may be perfectly symmetrical about some number c; that is, we may be able to pair off all the x's so that each x less than c can be paired with an x' such that c is exactly in the middle, between x and x'. In other words, the graph of the data is a mirror image of itself in the line drawn through the number c. In this case it is obvious that:

c is the mean of the numbers;
c is the median of the numbers;
c is the mid-point of the two extreme values

In such a case, if there is a single mode, then that mode is c; and if there are two modes, then their mean is c.

On the other hand, the geometric mean (for positive x's) is not equal to c.

If the data are not symmetrical, the five measures usually take different values. Almost nothing can be said about the necessary relationships between these five measures, except that the geometric mean must be less than or equal to the arithmetic mean — see the following exercise.

Exercise 1

Exercise 1
(2 minutes)

(i) Given any two positive numbers, x_1 and x_2, what can you say about the sign of $(x_1 - x_2)^2$?

(ii) Bearing in mind that

$$(x_1 - x_2)^2 = x_1^2 - 2x_1x_2 + x_2^2,$$

what can you deduce from (i) about

$$x_1^2 + 2x_1x_2 + x_2^2?$$

(iii) Deduce from your answer to (ii) that

$$(x_1 + x_2)^2 \geqslant 4x_1x_2,$$

and hence deduce that
arithmetic mean of $x_1, x_2 \geqslant$ geometric mean of x_1, x_2.

(iv) When are the arithmetic mean and the geometric mean equal? ■

Solution 16.4.1.1

Solution 16.4.1.1

(i) FALSE.
For example, 2 and 18 have geometric mean $\sqrt{2 \times 18} = 6$. Taking $k = 7$, $2 + 7$ and $18 + 7$ have geometric mean

$$\sqrt{9 \times 25} = 15$$
$$\neq 6 + 7.$$

(ii) TRUE.
geometric mean of y's

$$= \sqrt[n]{y_1 y_2 \cdots y_n}$$
$$= \sqrt[n]{a^n x_1 x_2 \cdots x_n}$$
$$= a \times (\text{geometric mean of } x\text{'s})$$

(iii) TRUE. (Note that $x_i \neq 0$ $(i = 1, 2, \ldots, n)$.)

geometric mean of y's

$$= \sqrt[n]{y_1 y_2 \quad y_n}$$
$$= \frac{1}{\sqrt[n]{x_1 x_2 \quad x_n}}$$
$$= \frac{1}{(\text{geometric mean of } x\text{'s})}.$$ ■

Solution 1

Solution 1

(i) $(x_1 - x_2)^2$ cannot be negative: so

$$(x_1 - x_2)^2 \geqslant 0.$$

(ii) $$0 \leqslant (x_1 - x_2)^2 = x_1^2 - 2x_1x_2 + x_2^2.$$

so that

$$x_1^2 + 2x_1x_2 + x_2^2 = x_1^2 - 2x_1x_2 + x_2^2 + 4x_1x_2$$
$$\geqslant 4x_1x_2.$$

(iii) $(x_1 + x_2)^2 = x_1^2 + 2x_1x_2 + x_2^2$
$$\geqslant 4x_1x_2,$$

so that, taking the positive square root of each side,

$$x_1 + x_2 \geqslant 2\sqrt{x_1x_2}.$$

Hence arithmetic mean

$$= \frac{x_1 + x_2}{2}$$
$$\geqslant \sqrt{x_1x_2} = \text{geometric mean}$$

i.e. arithmetic mean $\geqslant$ geometric mean.

(iv) If arithmetic mean = geometric mean, then

$$\frac{x_1 + x_2}{2} = \sqrt{x_1x_2}$$
$$(x_1 + x_2)^2 = 4x_1x_2$$
$$x_1^2 - 2x_1x_2 + x_2^2 = 0$$
$$(x_1 - x_2)^2 = 0$$
$$x_1 = x_2.$$ ■

16.4.3 Summary

Summarizing, there is more than one measure of central location, and they usually take different values for the same data unless the data are completely symmetrical. There is no universally "right" choice to make; you have to decide for yourself what you want to know, and why you want to know it. If you were a Government economist, you would be inclined to worry about the total purchasing power of the populace and would therefore think in terms of the mean earnings. If you were a prospective applicant for a job where the salary structure was pretty lop-sided, it would be sensible to consider medians. If you were a charlatan fortune teller, you might think in terms of the mode as this would maximize your chances of complete success. But these remarks do not give rules; they are merely guide lines — indications that there is no alternative to a correct appraisal by yourself. Get your own question crystal clear first, and then the matter is likely to settle itself. For purely mathematical reasons one would prefer to choose the measure which would be simple to manipulate in calculations — for example, the mean.

Acknowledgements

Grateful acknowledgement is made to the following sources for material used in this correspondence text:

Morton Kramer and The Royal Statistical Society for Table 4 from the article "Statistics of Mental Disorder in the U.S.A." by Morton Kramer, *Journal of the Royal Statistical Society*, Series A, vol. 132, Part 3 (1969).

The Controller of Her Majesty's Stationery Office for Tables 4, 26, 57, 277 from *Annual Abstracts of Statistics* (1953).

The Science Museum Library, photograph of Karl Pearson.

Unit No.	Title of Text
1	Functions
2	Errors and Accuracy
3	Operations and Morphisms
4	Finite Differences
5	NO TEXT
6	Inequalities
7	Sequences and Limits I
8	Computing I
9	Integration I
10	NO TEXT
11	Logic I — Boolean Algebra
12	Differentiation I
13	Integration II
14	Sequences and Limits II
15	Differentiation II
16	Probability and Statistics I
17	Logic II — Proof
18	Probability and Statistics II
19	Relations
20	Computing II
21	Probability and Statistics III
22	Linear Algebra I
23	Linear Algebra II
24	Differential Equations I
25	NO TEXT
26	Linear Algebra III
27	Complex Numbers I
28	Linear Algebra IV
29	Complex Numbers II
30	Groups I
31	Differential Equations II
32	NO TEXT
33	Groups II
34	Number Systems
35	Topology
36	Mathematical Structures